# OPX – A High Performance libfabric Provider for Omni-Path Networks

Dennis Dalessandro - Cornelis Networks

August, 2022

# Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH CORNELIS NETWORKS PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN CORNELIS NETWORKS'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, CORNELIS NETWORKS ASSUMES NO LIABILITY WHATSOEVER, AND CORNELIS NETWORKS DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF CORNELIS NETWORKS PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. CORNELIS NETWORKS PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Cornelis Networks may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Cornelis Networks reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice. Roadmap not reflective of exact launch granularity and timing. The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Any code names featured are used internally within Cornelis Networks to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Cornelis Networks to use code names in advertising, promotion or marketing of any product or services and any such use of Cornelis Networks' internal code names is at the sole risk of the user.

All products, computer systems, dates and figures specified are preliminary based on current expectations and are subject to change without notice. Material in this presentation is intended as product positioning and not approved end user messaging.

Performance tests are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Cornelis Networks technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration.

# Agenda

- Brief Intro
- Who/what is Cornelis Networks?
  - Doug will give much more detail in his talk
- What is OPX and why is it so great?

# Who is this guy?

- OSU Graduate
  - A long long time ago, aka 2004
- Former Researcher here at the Ohio Supercomputer Center
  - OSC-Springfield
  - Spent many days working from this very room (BALE)
- Now: Engineer for Cornelis Networks
  - Kernel developer
  - Maintain drivers upstream
  - Main distro point of contact
  - Work closely with user lib (OPX) developers

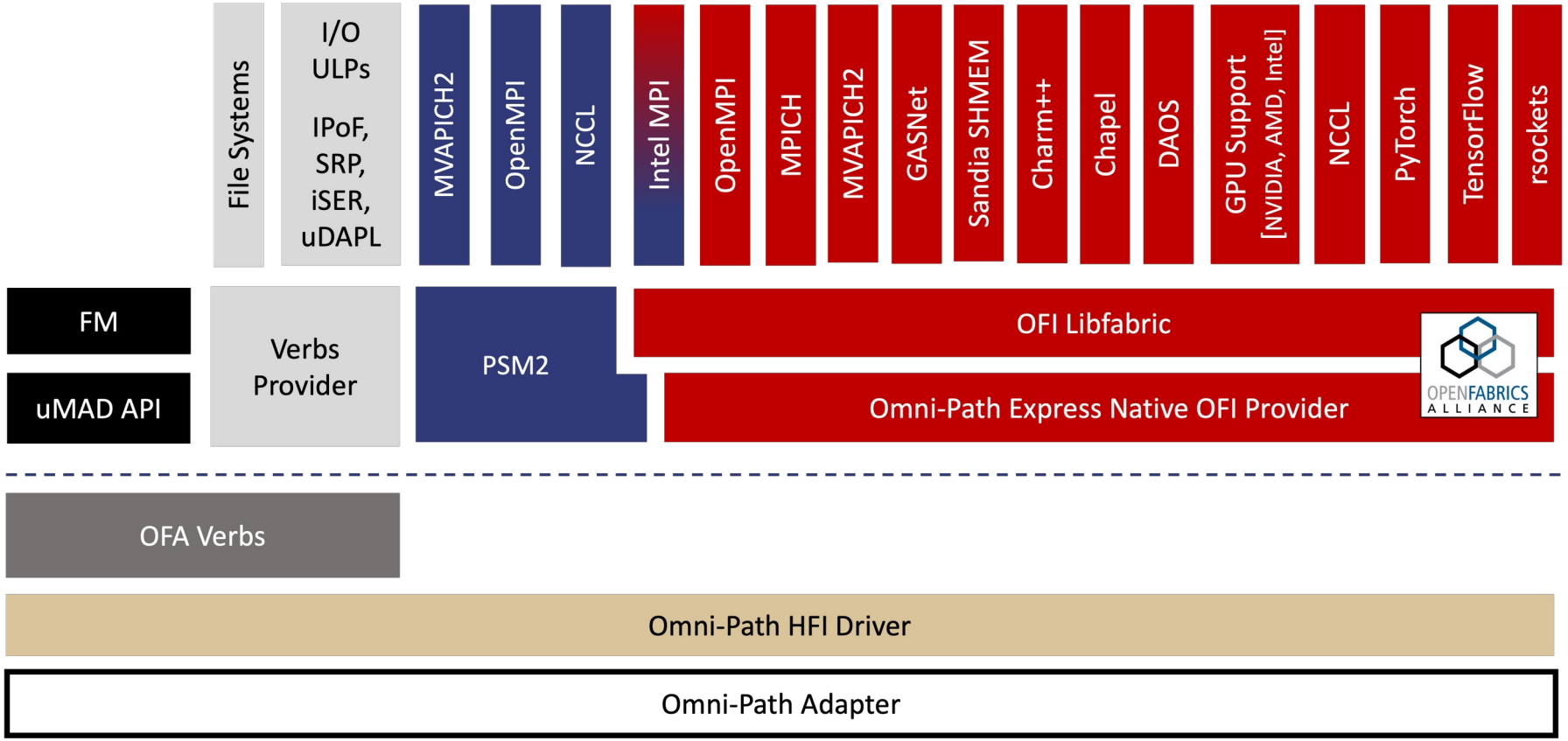
# Cornelis Networks

- From Startup -> QLogic (IB) -> Intel (OPA) -> Cornelis (OPX)
- Omni-Path Architecture (OPA) 100 Gbps Fabric
- Spun out of Intel 2+ years ago
  - Bring the customers and technology
  - Stand on our own now days
- Carry OPA flag forward and advance the technology
- Next Generation is 400Gbps (CN 5000) and beyond
  - See Doug Fuller’s talk

# What is OPX?

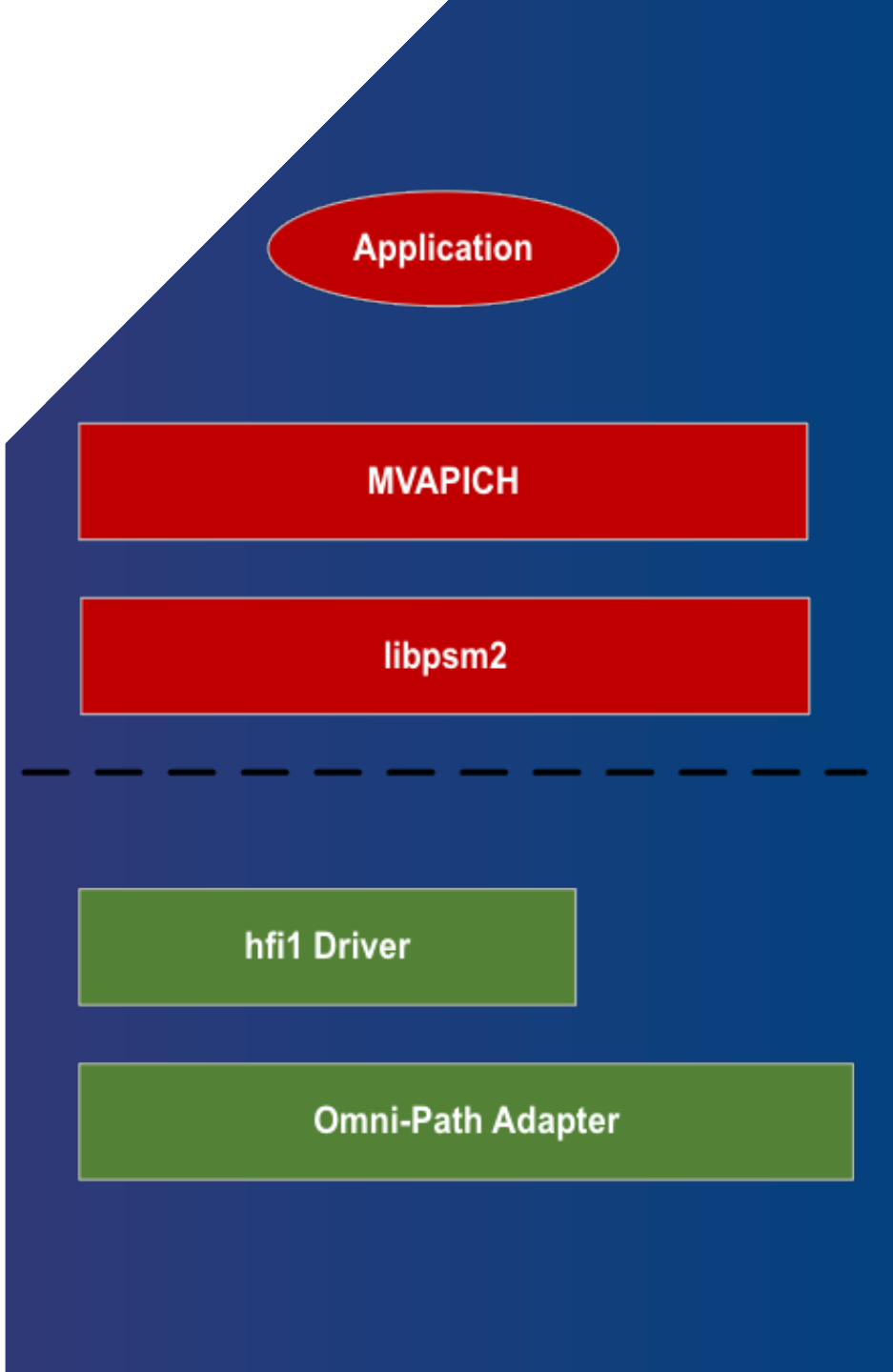
- OPX is a libfabric (OFI) provider
- Eventual goal is replacing PSM2
  - Not a re-write or refactor of PSM2
  - Origins in BGQ (Blue Gene) provider
- Highly optimized
  - Support coming for GPU
  - MVPACH2 w/libfabric now available
- Does not require changes to hfi1 driver or Fabric Manager
- Really is a drop in replacement

# Software Stack



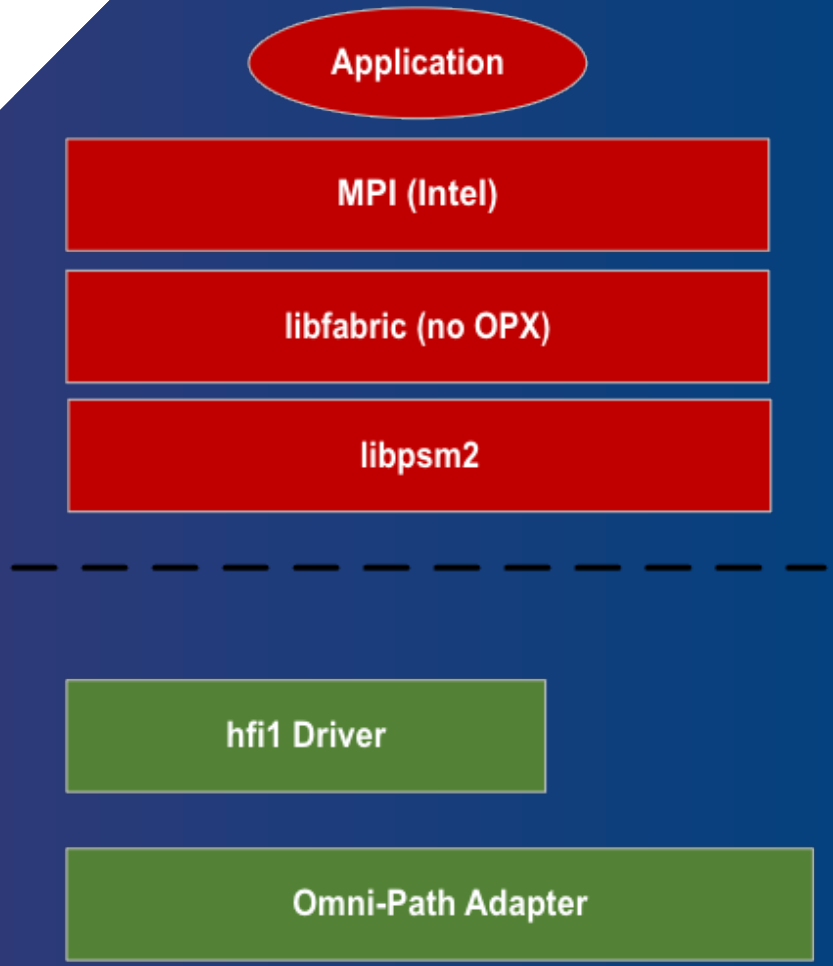
# MPI Direct to PSM2

- Pros
  - Library has direct access to PSM2
  - Fewer layers
- Cons
  - MPI Library has to support yet another API
  - PSM2 is not the path forward in OPA 400 and beyond



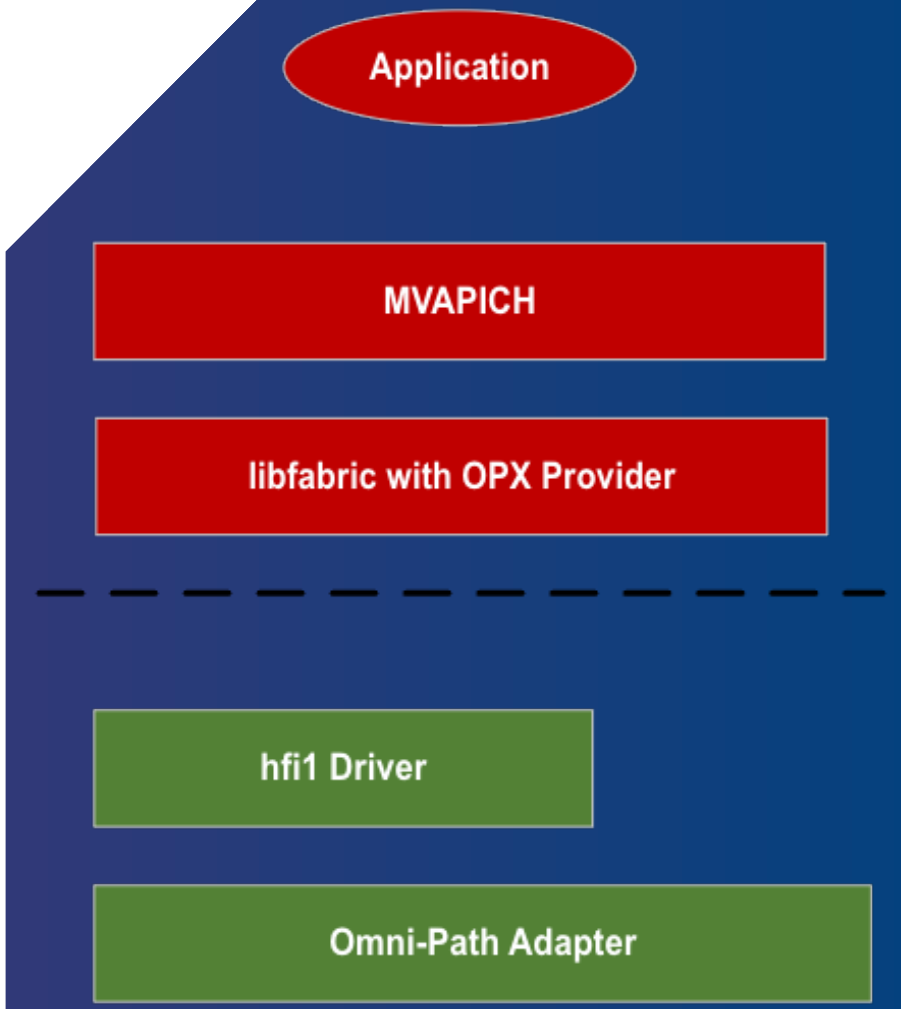
# MPI to Libfabric - pre OPX

- Pros
  - Allows MPI layer to code to a single API stack
- Cons
  - Not optimal code path
  - PSM2 is not a great semantic match to libfabric



# MPI to Libfabric with OPX

- Pros
  - Allows MPI layer to code to a single API
  - Allows MPI layer to have most efficient access to HW possible
  - Highest performance with a standard API
  - Same benefits as MPI->PSM2
  - Based on provider with proven performance
- Cons
  - New provider (under dev), takes time to upstream code
    - Upstream first is a key principle at Cornelis Networks

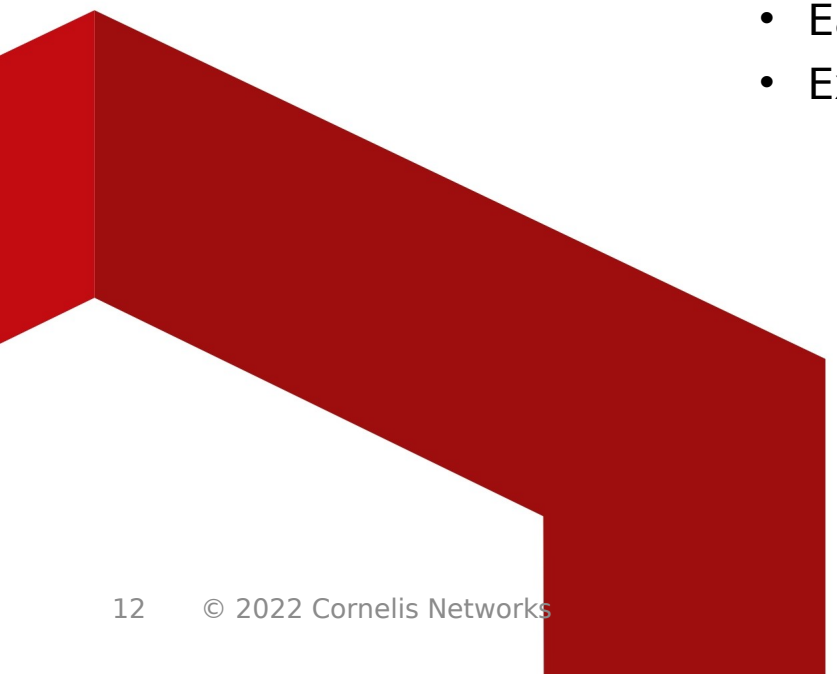


# OPX Kernel and HW Access

- Driver IOCTL
  - Create user contexts (end points)
  - Discover details about HW contexts
  - Register TID recv buffers
  - Misc other tasks not involving data movement
- Memory Map
  - Access to PIO buffers
  - Access to eager array and header queue
- Driver writev()
  - Send in a list of SDMA requests
  - Data is NOT written just details on where to find it
- Direct HW Access
  - PIO buffers are mapped to HW registers
  - Driver is NOT involved in I/O to the PIO space
- Kernel HW Access
  - SDMA Engines are programmed based on writev()

# Buffer Types Review

- Send
  - PIO – Programmed I/O (memory mapped) - Upstream Now
  - SDMA – DMA Engines on the hfi retrieve the user data – Upstream Soon
- Recv
  - Eager – Buffers that are filled as data comes in - Upstream Now
  - Expected – User buffers that are filled by the HW – Coming

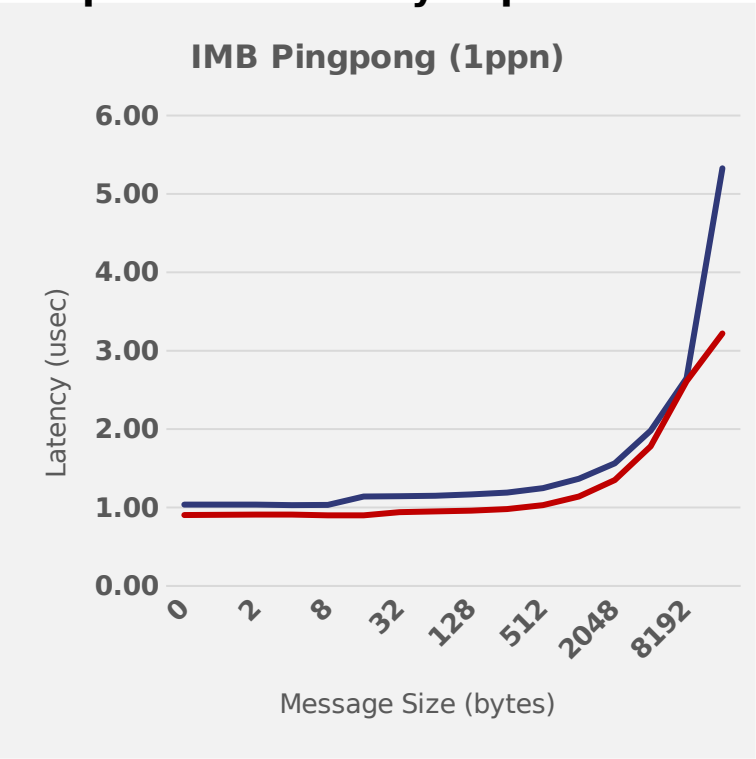


# Significant Performance Improvements

## Intel Xeon Icelake Platform

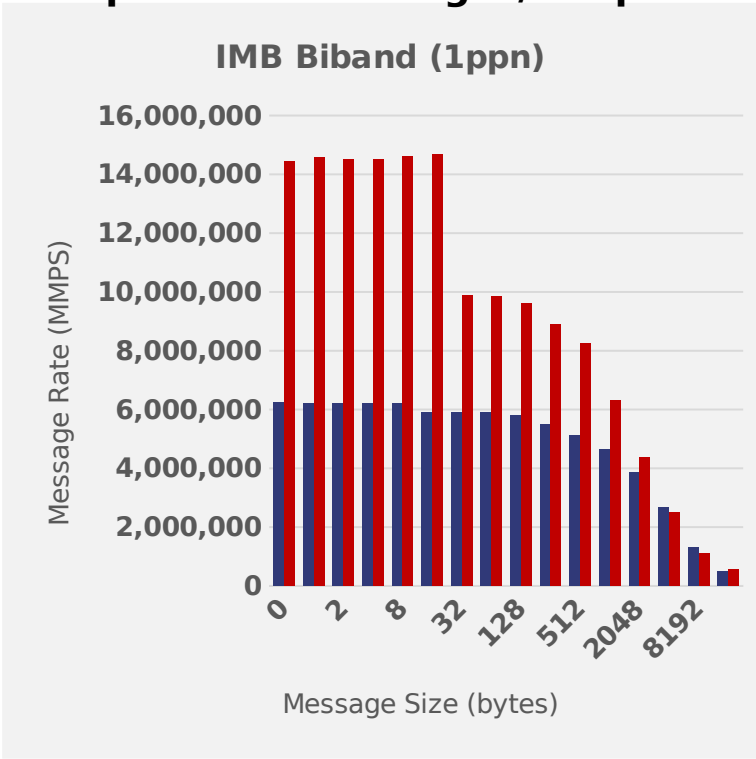
### Latency

**Up to 20% latency improvement**



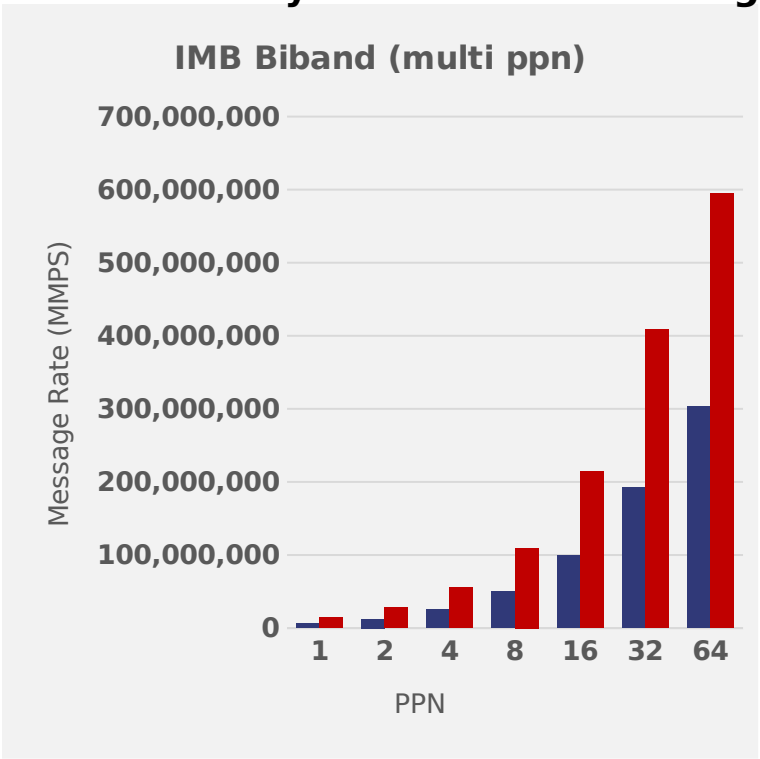

### Message Rate

**Up to 2.4X messages/sec per core**




### Scalability

**Linear Scalability at double the throughput**




PSM2 Provider  OPX Provider

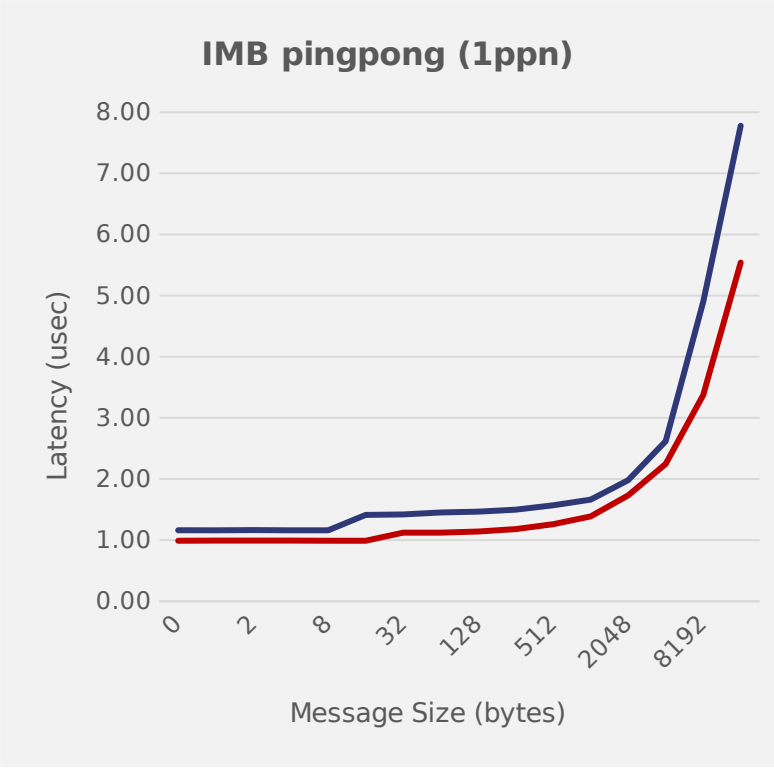
Test Configuration:
2-socket Intel® 3rd Generation Xeon® Scalable (Icelake) Platinum 8358, Dual Rail OPA100, BIOS: Snoop Hold-off Response Timer=11, Energy Efficient Turbo=DISABLED, C-States=DISABLED
Rocky Linux 8.4 (Green Obsidian), Kernel 4.18.0-305.19.1.el8_4.x86_64, IntelMPI 2019.6, IMB 2019.6, IFS 10.11.1.1.1, OPX Build 225

# Significant Performance Improvements

## AMD EPYC Milan Platform

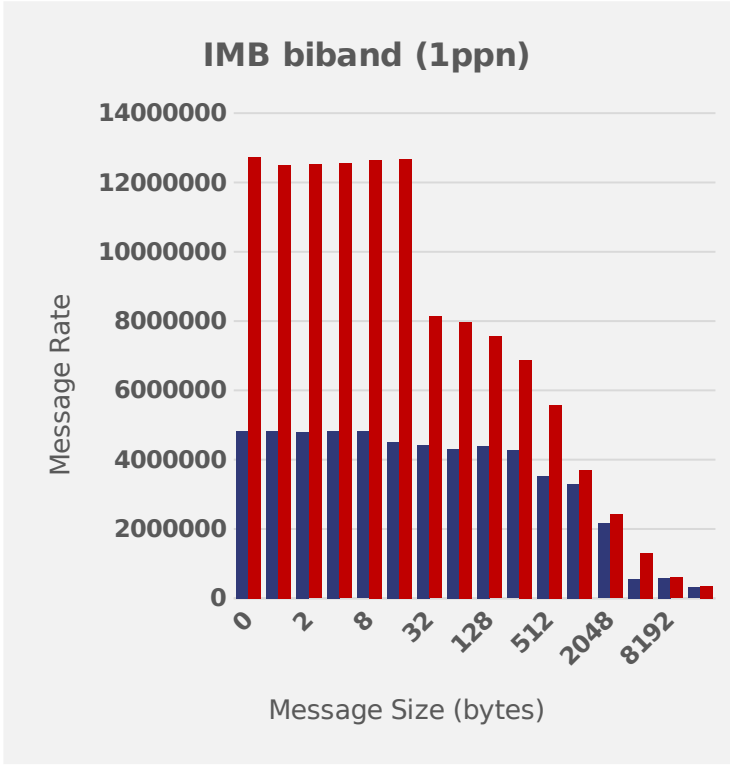
### Latency

**Up to 25% latency improvement**



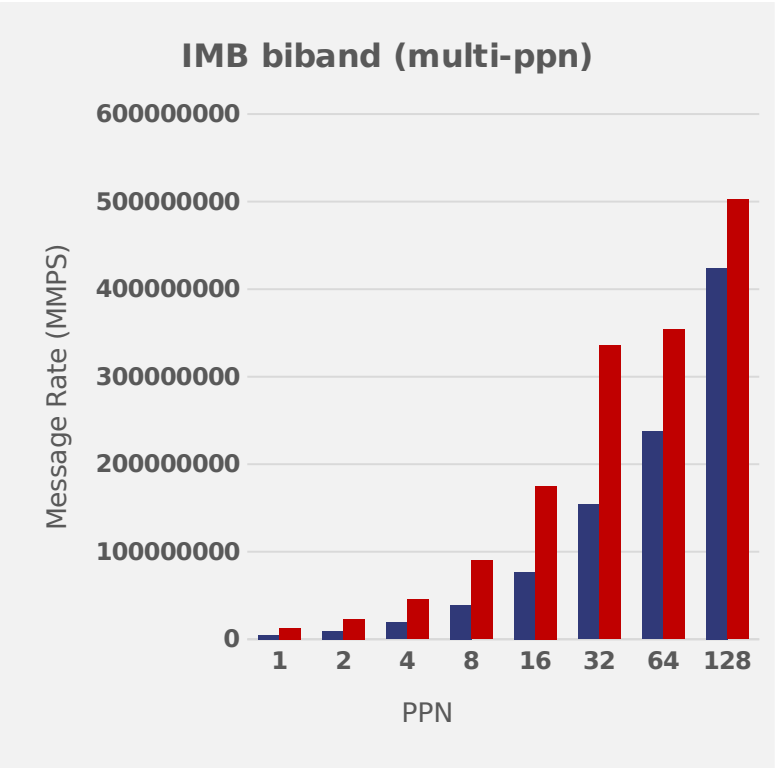

### Message Rate

**Up to 2.6X messages/sec per core**




### Scalability

**Linear Scalability at double the throughput**




PSM2 Provider  OPX Provider

Test Configuration:
2-socket AMD EPYC (Milan) 7713, Dual Rail OPA100, xGMI Frequency Locked, xGMI Link Width Locked, P-State Disabled, PCIe Slot Frequency Locked
CentOS Linux 8.3, IntelMPI 2019.6, IMB 2019.6, IFS 10.11.1.1.1, OPX Build 223

# Why is OPX better?

- Other than the universal application support…
- Semantic match for libfabric
- Optimized for performance
  - Textbook SW Engineering not as important as pure performance
  - Goal is performance, not making it easy on developers
  - Instruction count, cache line footprint matter
  - Focuses on libfabric where as PSM2 is a flexible application itself

# Example Improvement

- MPI_Recv()
  - Testing done with Intel SDE tool
    - Instruction count reduced by 38%
    - Cache lines for code reduced 60%
    - Cache line loads reduced 59%
    - New cache line access reduced 59%

# Current Status

- Early MVAPICH2 numbers look very promising!
- OPX Code is upstream as of v1.15 of Libfabric.
  - Focuses on small message/latency first
- Next Optimizations
  - Large message improvements – upstream release imminent
  - DAOS optimizations
  - GPU support
  - Full GA
- Available on GitHub, Distro*, OPXS Software Suite
  - Checkout Libfabric ‘main’ branch
- Get involved
  - Happy to take patches via GitHub

*Depends when a distro gets latest libfabric

# Thank You

www.cornelisnetworks.com